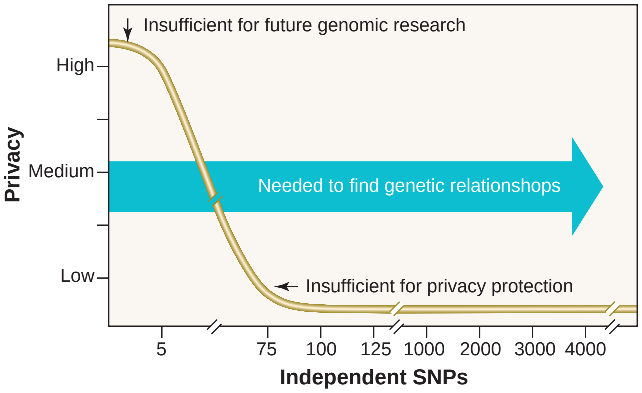
CORRECTED 3 SEPTEMBER 2004; SEE LAST PAGE

GENETICS

Genomic Research and Human Subject Privacy

Zhen Lin,[1] Art B. Owen,[2] Russ B. Altman[1]*

Interest in understanding how genetic variations influence heritable diseases and the response to medical treatments is intense. The academic community relies on the availability of public databases for the distribution of the DNA sequences and their variations. However, like other types of medical information, human genomic data are private, intimate, and sensitive. Genomic data have raised special concerns about discrimination, stigmatization, or loss of insurance or employment for individuals and their relatives (*1*, *2*). Public dissemination of these data poses nonintuitive privacy challenges.



Trade-offs between SNPs and privacy.

Unrelated persons differ in about 0.1% of the 3.2 billion bases in their genomes (*3*). Now, the most widely used forms of forensic identification rely on only 13 to 15 locations on the genome with variable repeats (*4, 5*). Single nucleotide polymorphisms (SNPs) contain information that can be used to identify individuals (*5*, *6*). If someone has access to individual genetic data and performs matches to public SNP data, a small set of SNPs could lead to successful matching and identification of the individual. In such a case, the rest of the genotypic, phenotypic, and other information linked to that individual in public records would also become available.

The world population is roughly 10^{10}. Specifying DNA sequence at only 30 to 80 statistically independent SNP positions will uniquely define a single person (*7*). Furthermore, if some of those positions have SNPs that are relatively rare, the number that need to be tested is much smaller. If information about kinship exists, a few positions will confirm it. Thus, the transition from *private* to *identifiable* is very rapid (see the figure).

Tension between the desire to protect privacy and the need to ensure access to scientific data has led to a search for new technologies. However, the hurdles may be greater than had been suspected. For example, one approach to protecting privacy is to limit the amount of high-quality data released and randomly to change a small percentage of SNPs for each subject in the database (*8*). Suppose that 10% of SNPs are randomly changed in a sequence of DNA, a fairly major obfuscation that would not please many genetics researchers. Our estimates (*7*) show that measuring as few as 75 statistically independent SNPs would define a small group that contained the real owner of the DNA. Disclosure control methods such as data suppression, data swapping, and adding noise would be unacceptable by similar arguments.

A second approach is to group SNPs into bins. Disregarding exact genomic locations of SNPs increases the number of records that share the same values, thus increasing confidentiality. Our calculations (*7*) show that such strategies do not protect privacy, because the pattern of binned values is unlikely to match anyone other than the owner of the DNA. Data analysis would be greatly complicated by binning, and the information content would be severely reduced or even eliminated.

Until technological innovations appear, solutions in policy and regulations must be found. We are building the Pharmacogenetics and Pharmacogenomics Knowledge Base (*8*, *9*), which contains individual genotype data and associated phenotype information. No genetic data will be provided unless a user can demonstrate that he or she is associated with a bona fide academic, industrial, or governmental research unit and agrees to our usage policies (including audit of data access) (*10*). Although this does not prevent data abuse, it provides a way to monitor usage.

Social concerns about privacy are intricately connected to beliefs about benefits of research and trustworthiness of researchers and governmental agencies. In the United States, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the associated Privacy Rules of 2003 (*11*) generally forbid sharing identifiable data without patient consent. However, they do not specifically address use or disclosure policies for human genetic data. Recent debates in Iceland, Estonia, Britain, and elsewhere (*12–15*), reveal a range of views on the threats posed by genetic information. The United States may be at one end of this spectrum, as its citizens seem to strongly desire health privacy. Whatever the setting, we recommend explicit clarifications to rules and legislation (such as HIPAA), so that they explicitly protect genetic privacy and set strong penalties for violations. These clarifications should define entities authorized to use and exchange human genetic data and for what purposes.

[1]Department of Genetics, Stanford University School of Medicine, CA 94305–5120, USA. [2]Department of Statistics, Stanford University, CA 94035–4065, USA.

*To whom correspondence should be addressed. E-mail: russ.altman@stanford.edu

References and Notes

1. M. R. Anderlik, M. A. Rothstein, *Annu. Rev. Genomics Hum. Genet.* **2**, 401 (2001).
2. P. Sankar, *Annu. Rev. Med.* **54**, 393 (2003).
3. W. H. Li, L. A. Sadler, *Genetics* **129**,513 (1991).
4. L. Carey, L. Mitnik, *Electrophoresis* **23**, 1386 (2002).
5. H. D. Cash *et al.*, *Pac. Symp. Biocomput.* **2003**, 638 (2003).
6. National Commission on the Future of DNA Evidence, *The Future of Forensic DNA Testing: Predictions of the Research and Development Working Group* (National Institute of Justice, U.S. Department of Justice, Washington, DC, 2000).
7. See supporting online material for further discussion.
8. L. C. R. J. Willenborg, T. D. Waal *Elements of Statistical Disclosure* Control (Springer, New York, 2001).
9. T. E. Klein *et al.*, *Pharmacogenomics J.* **1**, 167 (2001).
10. www.pharmgkb.org/home/policies/index.jsp
11. *Fed. Regist.* **67**, 53181 (2002).
12. R. Chadwick, *BMJ* **319**, 441 (1999).
13. L. Frank, *Science* **290**, 31 (2000).
14. M. A. Austin *et al.*, *Genet. Med.* **5**, 451 (2003).
15. V. Barbour, *Lancet* **361**, 1734 (2003).
16. Supported in part by NIH/NLM Biomedical Informatics Training Grant LM007033 (Z.L.), NSF Grant DMS-0306612 (A.B.O.), and the NIH/NIGMS Pharmacogenetics Research Network and Database U01-GM61374 (R.B.A). We thank J. T. Chang, B. T. Naughton, T. E. Klein, and reviewers.

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5681/183/DC1

ERRATUM

post date 3 September 2004

Policy Forum: "Genomic research and human subject privacy" by Z. Lin *et al.* (9 July 2004, p. 183). In the figure, the word on the colored arrow should be "relationships."